INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation

Chenda Li, Yanmin Qian [*]

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai
lichenda1996@sjtu.edu.cn, yanminqian@sjtu.edu.cn

Abstract

Solving the cocktail party problem with the multi-modal approach has become popular in recent years. Humans can focus on the speech that they are interested in for the multi-talker mixed speech, by hearing the mixed speech, watching the speaker, and understanding the context what the speaker is talking about. In this paper, we try to solve the speaker-independent speech separation problem with all three audio-visual-contextual modalities at the first time, and those are hearing speech, watching speaker and understanding contextual language. Compared to the previous methods applying pure audio modal or audio-visual modals, a specific model is further designed to extract contextual language information for all target speakers directly from the speech mixture. Then these extracted contextual knowledge are further incorporated into the multi-modal based speech separation architecture with an appropriate attention mechanism. The experiments show that a significant performance improvement can be observed with the newly proposed audio-visual-contextual speech separation.

Index Terms: speech separation, audio-visual, multi-modal, contextual language embedding

1. Introduction

When conversations take place in a complex environment, it is not difficult for humans to focus on and understand the speech that they are interested in from the multi-talker mixed speech. However, there still remains a huge challenge for machines to solve that problem like humans. This problem is defined as cocktail party problem [1, 2, 3].

The speech separation technology is one of the key points in solving the cocktail party problem. In recent years, combining the deep leaning methods, great progress has been made in speech separation. When performing speech separation with deep learning approaches, one important thing that needs to be solved is the label permutation problem. Deep Clustering (DPCL) [4] solves this problem by projecting the mixed speech into high level embeddings, and using clustering algorithm to perform the separation. Furthermore, the permutation invariant training (PIT) [5, 6, 7] is proposed, and it is a simple and effective algorithm for tackling the label permutation problem.

In real scenarios, the information that can be utilized by humans for speech separation is more than speech itself. For example, people pay subconscious attention to the vision of the speakers, e.g. the speaker's position or lip movements, when the interfered speech or background noise becomes too strong. Inspired by this human mechanism, recent researches [8, 9, 10, 11, 12] have introduced the visual modal into the traditional audio-only speech separation. Different strategies of utilizing the visual clues in speech separation have been explored. In [8, 9], the visual clue is converted into visual representations, and concatenated with the audio representation frame by frame. In [10, 11], more robust approaches have been explored to deal with the situation that the visual clues are temporarily absent. In our previous work [12], an attention mechanism has been developed to make better use of the visual clues. By taking the both advantages of the audio and visual knowledge, the performance of speech separation has been significantly improved.

It inspires us to explore other more information in addition to audio and visual ones. In a complex environment, besides focusing on the modal of speech and vision of the speaker, humans will also utilize the contextual language understanding to address the multi-talker mixed speech, i.e. what the speaker has talked about at that time. It is believed that the modal of contextual language information in the speech should also help in speech separation. However the extraction of the contextual language information is not straightforward from the multi-talker mixed speech. We need to explore how to extract the contextual information from the mixed speech accurately and then how to incorporate the contextual language modal in speech separation.

The recent works in [13, 14] have the similar idea in speech enhancement and speech separation, but only with relative simple attempts. A two-stage approach is used in [14], and the phonetic information is only used in the second stage. In that method, the first step is a conventional speech separation that does not utilize the target speaker's contextual information. Then, they extract the contextual information from the separated speech, and the second speech separation stage is performed with that contextual information to achieve better performance.

In this paper, we proposed a better approach to introduce contextual language modal into audio-visual speech separation, and we named it audio-visual-contextual speech separation. Firstly, oracle contextual language embedding extracted from the target speech is incorporated, which is proven helpful in speech separation; Secondly, a model that directly extracts contextual information from the mixed speech is designed and constructed. Finally, a strategy to incorporate the predicted contextual language modal has been explored. With the proposed method, our new model achieves a large improvement compared to the baseline.

The rest of this paper is organized as follows. Section 2 describes the basic audio-visual speech separation; In Section 3, the proposed audio-visual-contextual architecture using all three modals is described in detail, including hearing speech, watching speaker and understanding contextual language. The experimental results, comparison and analysis are given and dis-

[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-2028

cussed in Section 4. Finally, Section 5 concludes the paper.

2. Audio-Visual Speech Separation

The audio-visual architecture proposed in our previous work [12] is firstly revisited in this section. It is an extension from the work in [8], with both visual streams in the mixed speech, and achieves a better performance [12]. The separation of the speech mixture is carried out in the Time-Frequency (T-F) domain. Considering a linearly mixed speech of two speakers A and B:

$$y(n) = \sum_{s \in A,B} x_s(n)\,, \tag{1}$$

after Short-Time Fourier Transformation (STFT), the signal in T-F domain can be written as:

$$Y(t,f) = \sum_{s \in A,B} X_s(t,f) \tag{2}$$

Let $\mathbf{y}_{s,i} = [Y(i,1), Y(i,2), \cdots, Y(i, \frac{N}{2}+1)]^T \in \mathbb{C}^{\frac{N}{2}+1}$ denote a singe frame of the mixed STFT, where N is the size of STFT. The mixed STFT of T frames can be written as: $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T] \in \mathbb{C}^{(\frac{N}{2}+1)\times T}$. The magnitude spectrum of $\mathbf{Y}$ can be denoted as $|\mathbf{Y}| = [|\mathbf{y}_1|, |\mathbf{y}_2|, \cdots, |\mathbf{y}_T|] \in \mathbb{R}^{(\frac{N}{2}+1)\times T}$. Then, denote the visual representation of the two target speaker A and B as $\mathbf{V}_A, \mathbf{V}_B \in \mathbb{R}^{D \times \frac{T}{4}}$, where D is the number of dimension of each frame. Our audio-visual speech separation network Net can be abstractly expressed as:

$$M_A, M_B = Net(|\mathbf{Y}|, \mathbf{V}_A, \mathbf{V}_B) \tag{3}$$

where M_A, M_B are the estimated magnitude mask [15].

As Figure 1 shows, the audio-visual separation network takes the speech magnitude spectrum of the mix speech $|\mathbf{Y}|$, the corresponding visual representations of two speakers $\mathbf{V}_A$ and $\mathbf{V}_B$ as the input. The input representations are processed by different 1-D $ResNets$ [16]. Each $ResNet$ consists of a stack of basic blocks, and each basic block contains a 1-D convolution layer with residual connection, a ReLU activation layer and a batch normalization layer. Some of the basic blocks contain an extra up-sampling or down-sampling layer. The visual representations $\mathbf{V}_A$ and $\mathbf{V}_B$ are firstly processed by a shared weight $ResNet_V$, to get the high level visual representations $\mathbf{V}_A^R$ and $\mathbf{V}_B^R$. The magnitude spectrum of the mix speech $|\mathbf{Y}|$ is processed by $ResNet_M$ to get the high level audio representation $\mathbf{Y}^R$. There are 2 down-sampling layers in $ResNet_M$ with a down-sampling factor of 2, since in our setup, every visual representation frame corresponds to 4 audio frames. The high-level representations are then concatenated over channels to get a fusion representation $\mathbf{F} = [\mathbf{V}_A^R; \mathbf{V}_B^R; \mathbf{Y}^R]$. The fusion representation is passed to $ResNet_{F_A}$ and $ResNet_{F_B}$, and then activated by sigmoid to estimate magnitude masks M_A and M_B. The estimated magnitude masks are applied to the mixed magnitude spectrum by element-wise multiplication to obtain the predicted magnitude spectrum:

$$\begin{aligned} |\mathbf{X}_A^*| &= |\mathbf{Y}| \odot M_A \\ |\mathbf{X}_B^*| &= |\mathbf{Y}| \odot M_B \end{aligned} \tag{4}$$

The $L1$ loss is used in training, and the optimization objective is:

$$\mathcal{L}_\alpha = \frac{\|(|\mathbf{X}_A| - |\mathbf{X}_A^*|)\|_1 + \|(|\mathbf{X}_B| - |\mathbf{X}_B^*|)\|_1}{2} \tag{5}$$

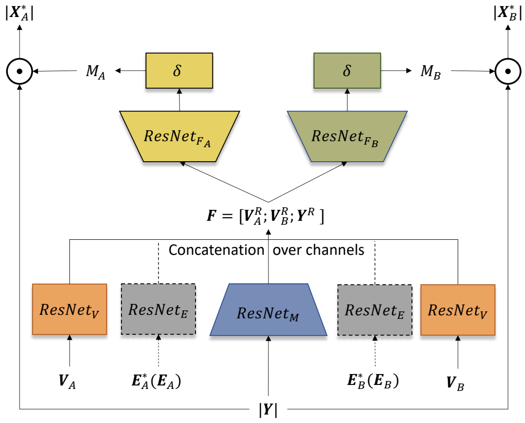

Figure 1: *Audio-Visual(-Contextual) speech separation architecture. The $ResNets$ with the same color are weight-shared. Noted that the $ResNet_E$ is the newly proposed contextual language embedding modal which is described in Section 3.*

where $|\mathbf{X}|_A$ and $|\mathbf{X}|_B$ are the target magnitude spectrum of two speakers in the mixed speech, respectively.

In the stage of separation, the estimated magnitude spectrum and the phase spectrum of $\mathbf{Y}$ are used to reconstruct the predicted STFT spectrum, then the predicted speech can be recovered by inverse Short-Time Fourier transform (iSTFT).

3. Audio-Visual-Contextual Speech Separation

In addition to the visual modality, we further explore the contextual language modal for speech separation.

3.1. Contextual Language Embedding Learning

In attention based end-to-end speech recognition models [17, 18, 19], the encoder is considered to encode the contextual information of the speech signal. The work in [14] has proven that explicitly incorporating the contextual information, including the phonetic and linguistic information of each speaker, can help in boosting the performance of speech separation. However it is a two-stage approach. The first step is a normal speech separation without using the contextual information. The contextual information is extracted from the separated speech in the first stage, and then the second separation with contextual information is constructed. This method has some constraints: it highly relies on the performance of the first stage separation module which influences accuracy of the contextual information extraction, and in the other hand usually the clean speech from the target speaker is also not available in real scenarios.

Here, we propose a more direct and effective method to extract the contextual language embedding and further integrate it with audio and visual modals into speech separation. As shown in Figure 2, the whole framework of the proposed contextual language embedding learning is illustrated. First a joint CTC-attention based end-to-end single speaker speech recognition model [19] is firstly well trained with single-speaker data using the ESPnet [20] toolkit . Using this pre-trained single-speaker ASR model, the encoder can generate the oracle contextual language embeddings $\mathbf{E}_A$ and $\mathbf{E}_B$ for the two mixed speakers A

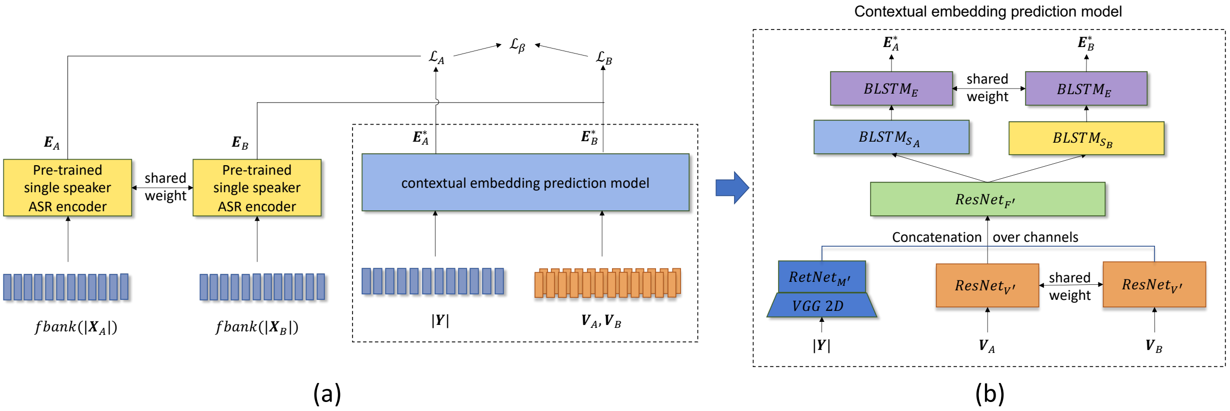

Figure 2: *The architecture of proposed contextual language embedding learning. The modules with the same color are weight-shared. (a) The whole architecture of contextual language embedding learning; (b) The structure of the contextual embedding prediction model.*

and B. These oracle contextual embeddings $\mathbf{E}_A$ and $\mathbf{E}_B$ can be used in later separation module training directly, or can be further used as the labels to train the contextual embedding prediction module.

In the contextual prediction model, both the spectrum feature of the mixed speech and two speakers' visual representations are utilized as the inputs. As Figure 2 shows, the visual representations are processed by a shared weight 1-D $ResNet_{V'}$, and the mixed magnitude spectrum $|\mathbf{Y}|$ is processed by a 2-D VGG-like [21] layer and a 1-D $ResNet_{M'}$. Then, the high-level representations are concatenated into a fusion representation. The fusion representation is then processed by an 1-D $ResNet_{F'}$. We use two separated bidirectional long short-term memory (BLSTM) layers, i.e. $BLSTM_{S_A}$ and $BLSTM_{S_B}$, and a shared encoder BLSTM layer $BLSTM_E$ for each speaker to predict contextual embeddings for individual speaker, and the generated $\mathbf{E}_A^*$ and $\mathbf{E}_B^*$ are predicted contextual embedding for both speakers in the mixed speech.

The training criterion can be written as:

$$\mathcal{L}_\beta = \frac{1}{2}(\|\mathbf{E}_A - \mathbf{E}_A^*\|_2 + \|\mathbf{E}_B - \mathbf{E}_B^*\|_2) \tag{6}$$

3.2. Audio-Visual-Contextual speech separation

The predicted (or oracle) contextual language embeddings can be then integrated with audio and visual modals to construct audio-visual-contextual speech separation as illustrated in Figure 1. A shared weight $ResNet_E$ is added and it transforms the contextual embeddings $\mathbf{E}_A^*(\mathbf{E}_A)$ and $\mathbf{E}_B^*(\mathbf{E}_B)$ into high-level representations $\mathbf{E}_A^{*R}$ and $\mathbf{E}_B^{*R}$ for speech separation. Then, similar to the audio-visual system, all the high-level representations are concatenated together as a fusion representation $\mathbf{F} = [\mathbf{V}_A^R; \mathbf{V}_B^R; \mathbf{Y}^R; \mathbf{E}_A^{*R}; \mathbf{E}_B^{*R}]$. The following pipeline is the same as the model introduced in Section 2.

3.3. Attention with Multi-Modal Embeddings

An attention mechanism [22, 23] is developed to better utilize multi-modal information in our proposed audio-visual-contextual speech separation system.

Before the fusion step described in Section 3.2, the high-level representations $\mathbf{V}_A^R$ and $\mathbf{E}_A^{*R}$ are first concatenated together, and projected by a shallow net $ResNet_{VE}$ to get a fusion representation $\mathbf{C}_A$. The same procedure for speaker B to get $\mathbf{C}_B$. $\mathbf{C}_A$ and $\mathbf{C}_B$ can be considered as the clue information for target speakers.

The scaled dot product attention score matrix $\mathbf{A}$ is computed between $\mathbf{C}_A$ and $\mathbf{C}_B$:

$$\mathbf{A}_{i,j} = \frac{\mathbf{C}_A[:,i] \cdot \mathbf{C}_B[:,j]}{\sqrt{D}} \tag{7}$$

where D is the dimension number of $\mathbf{C}_A$ and $\mathbf{C}_B$. Then, the attention score matrix $\mathbf{A}$ is converted into attention features with a learnable fully connected layer $\mathbf{W}$:

$$\begin{aligned} \mathbf{\Gamma}_A &= \mathbf{W} \cdot \mathbf{A}^T \\ \mathbf{\Gamma}_B &= \mathbf{W} \cdot \mathbf{A} \end{aligned} \tag{8}$$

$\mathbf{W}$ projects $\mathbf{A} \in \mathbb{R}^{L \times L}$ into $\mathbf{\Gamma} \in \mathbb{R}^{D \times L}$, where L is the max frame length in the dataset. Padding positions of $\mathbf{\Gamma}_A$ and $\mathbf{\Gamma}_B$ are masked in the implementation. Finally, all high-level representations are combined together, $\mathbf{F} = [\mathbf{C}_A; \mathbf{C}_B; \mathbf{Y}^R; \mathbf{\Gamma}_A; \mathbf{\Gamma}_B]$.

4. Experiments

4.1. Data Preparation

The speech separation model and the contextual embedding prediction model are trained on LRS2 [24] dataset. It is an audio-visual dataset collected from BBC television. We also use the LibriSpeech corpus [25] in end-to-end single-speaker ASR training.

Visual representation: We use a pre-trained lip reading net described in [26, 27] to extract visual representations from LRS2 dataset. For each frame of a video, face region of the speaker is firstly cropped, and then processed by the pre-trained model to generated a 512-dimensional feature.

Audio representation: In LRS2 dataset, the audio is recorded at a sample rate of 16kHz, and the frame rate of the video is 25fps. As for STFT, the windows size is set to 40ms and the hop length is 10ms. With this setup, each frame of the magnitude spectrum is 321-dimensional, and every 4 frames of the magnitude spectrum correspond to one single frame of the visual representation.

Contextual learning: In end-to-end single-speaker ASR training, the input features is converted to 80 dimensional log-mel filterbank coefficients. The predicted or oracle contextual embedding is 512-dimensional. The ASR encoder performs 4-time

Table 1: *The details of ResNets in Figure 1 and Figure 2.* ***N****: The number of residual blocks.* ***C****: Number of convolution channels;* ***O****: The output dimensional, if different from* ***C****, an extra projection layer is included;* ***K****: Kernel size;* ***D/U****: downsampling or upsampling factor on the time scale.*

$ResNets$	N	C	O	K	U/D
V	10	1024	1024	5	-
M	5	1024	1024	5	D: 4×
F_A	15	1024	321	5	U: 4×
F_B	15	1024	321	5	U: 4×
E	3	1024	1024	5	-
$V^{'}$	5	1024	1024	5	-
$M^{'}$	3	1024	1024	5	-
$F^{'}$	5	1024	1024	5	-

subsampling over time scale of the input features. So the oracle contextual embedding have the same length as the visual representation.

Synthetic Audio: The mixed audio is generated from two target audios randomly picked from the LRS2 dataset. The target audios are linearly mixed, where the shorter audio is padded to the same length as the longer one.

4.2. Network Configuration and Training Details

The joint CTC/attention based end-to-end single-speaker ASR model is trained with the LibriSpeech 960h corpus. The training procedure follows the recipe in ESPnet toolkit [20]. After convergence on the LibriSpeech dataset, the model is then finetuned with the LRS2 training set. The final well-trained ASR models reaches 8.2% word error rate (WER) on LRS2 test set. The ASR encoder used to extract oracle ASR features is a 5-layers BSTLM with projection, each layer containing 512 units, and the encoder performs 4-time subsampling over the time scale.

The VGG-like in contextual embedding prediction model (Figure 2) contains 4 layers of 2-D convolution. In each convolution, the kernel size is 3, and the channel number of convolution layers is 64-64-128-128. Two max pooling layers are included in the VGG-like block, which perform 4-times subsampling in the time scale. The separated BLSTM networks consist of 2 layers with 512 units, and the shared-weight BLSTM encoder consists of 1 layer with 512 units. The dropout rate of BSLTMs is set to 0.2. The details of ResNets in the contextual embedding prediction model are listed in Table 1. The Adam optimizer with weight decay 10^{-6} is used during training. The learning rate is initially set to 10^{-4}, and then is reduced by the factor 0.7 in every 3 epochs. The batch size is set to 16, and 4 GTX-2080Ti GPUs are used for data parallel training.

The details of $ResNets$ in the audio-visual or audio-visual-contextual speech separation networks are listed in Table 1. The training procedure is almost the same as that in our previous work [12], except for the data length. In order to maintain the consistency of the context information, in this paper, the input data is not clipped to a fixed length. 4 GTX-2080Ti GPUs are used for data parallel training of the speech separation model, and the batch size is set to 32.

4.3. Results and Analysis

We adopt the signal-to-distortion-ration (SDR) [28], short-time objective intelligibility (STOI) [29] and perceptual evaluation of speech quality score (PESQ) [30] as evaluation metrics.

Table 2: *Results comparison of the Audio-Visual-Contextual Speech Separation models.* ***GT****: ground truth phase;* ***MX****: noisy phase;* ***C-tr****: contextual embedding in model training;* ***C-tt****: contextual embedding used in testing;* ***O****: oracle contextual embedding;* ***P****: predicted contextual embedding.*

C-tr	C-tt	Φ	SDR	STOI	PESQ
-	-	MX	11.18	0.730	3.09
O	O	MX	11.68	0.944	3.28
P	P	MX	11.51	0.937	3.12
P + O	P	MX	11.76	0.940	3.23
-	-	GT	16.82	0.952	3.50
O	O	GT	18.31	0.967	3.76
P	P	GT	17.77	0.961	3.66
P + O	P	GT	18.17	0.964	3.70

To evaluate the upper bound of incorporating contextual embedding, the oracle contextual embedding is firstly used in both training and evaluation. As Table 2 shows, the speech separation system with oracle contextual embedding shows large improvements over the audio-visual speech separation system on all the metrics. We then evaluate the new audio-visual-contextual model with the predicted contextual embedding, since the oracle contextual embedding is actually not available in real application. Different contextual embeddings usages in training and testing are compared and listed in Table 2. The experimental results show that the contextual embedding extracted with the proposed model can also significantly improves the speech separation upon the strong audio-visual two-modal system. We further evaluate the proposed multi-modal attention mechanism described in Section 3.3, and the results are illustrated in Table 3. It is observed that an additional and consistent improvement can be obtained using the proposed attention with multi-modal embeddings.

Table 3: *Results of combining the proposed attention mechanism.* ***P-P****: the model trained and evaluated both with predicted contextual embedding;* ***GT****: ground truth phase;* ***MX****: noisy phase.*

Model	GT			MX		
	SDR	PESQ	STOI	SDR	PESQ	STOI
P-P	17.77	3.66	0.961	11.51	3.12	0.937
+ att	18.12	3.68	0.964	11.68	3.22	0.940

5. Conclusions

In this paper, we proposed a new multi-modal speech separation architecture, including audio-visual-contextual three modalities. A specific model extracting contextual language information directly from multi-talker mixed speech has been designed, and these contextual language knowledge is further incorporated with other modals with an appropriate attention mechanism to perform speech separation. With the proposed audio-visual-contextual architecture, we can obtain a significant improvement for speech separation.

6. Acknowledgements

This work has been supported by the China NSFC project No. U1736202. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. H. McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[3] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*. IEEE, 2016, pp. 31–35.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[6] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[7] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *Proc. ISCA Interspeech*, pp. 3244–3248, 2018.

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 112, 2018.

[10] T. Afouras, J. S. Chung, and A. Zisserman, "My Lips Are Concealed: Audio-Visual Speech Enhancement Through Obstructions," pp. 4295–4299, 2019.

[11] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues," *Proc. ISCA Interspeech*, pp. 2718–2722, 2019.

[12] C. Li and Y. Qian, "Deep audio-visual speech separation with attention mechanism," in *Proc. IEEE ICASSP*, 2020, pp. 7314–7318.

[13] B. Wu, M. Yu, L. Chen, M. Jin, D. Su, and D. Yu, "Improving speech enhancement with phonetic embedding features," in *Proc. IEEE ASRU*. IEEE, 2019, pp. 645–651.

[14] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy, and Y. Mitsufuji, "Improving voice separation by incorporating end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2020, pp. 41–45.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. ASLP.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*. IEEE, 2016, pp. 4960–4964.

[18] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[19] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[20] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. ISCA Interspeech*, 2018, pp. 2207–2211.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. IEEE ICASSP*. IEEE, 2015, pp. 5206–5210.

[26] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Proc. ISCA Interspeech*, 2017, pp. 3652–3656. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-85

[27] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proc. IEEE ICASSP*. IEEE, 2018, pp. 6548–6552.

[28] C. Févotte, R. Gribonval, and E. Vincent, "Bss-eval toolbox user guide : Revision 2.0," IRISA, Tech. Rep. 1706, Apr. 2005. [Online]. Available: https://www.irit.fr/~Cedric.Fevotte/publications/techreps/BSSEVAL2userguide.pdf

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.